

Probabilidade e Estatística

Aula 2

Apresentação de dados em Tabelas e Gráficos

Leitura: Levine et al. Capítulo 2



Objetivos

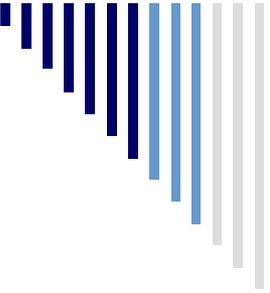
Nesta parte, vamos aprender:

- A conhecer as fontes e tipos de dados
- A desenvolver tabelas e gráficos para conjuntos de dados
- Os princípios de apresentação adequada de gráficos



Fontes de Dados

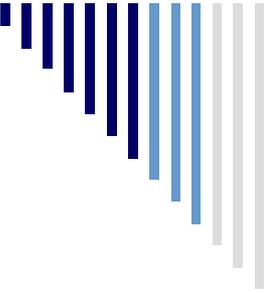
- A estatística é a ciência de análise de dados que representam uma situação onde existe incerteza.
- Mas de onde vem os dados normalmente utilizados?
 - 1) Fontes Primárias: os dados são coletados pelo pesquisador
 - dados de uma pesquisa política
 - dados coletados por um experimento*
 - 2) Fontes Secundárias: a pessoa faz a análise de dados coletados por terceiros.
 - dados do censo (ibge, inep,...)
 - dados de revistas/jornais impressos ou publicados na internet



Fontes de Dados

Os dados coletados (ou amostras) devem:

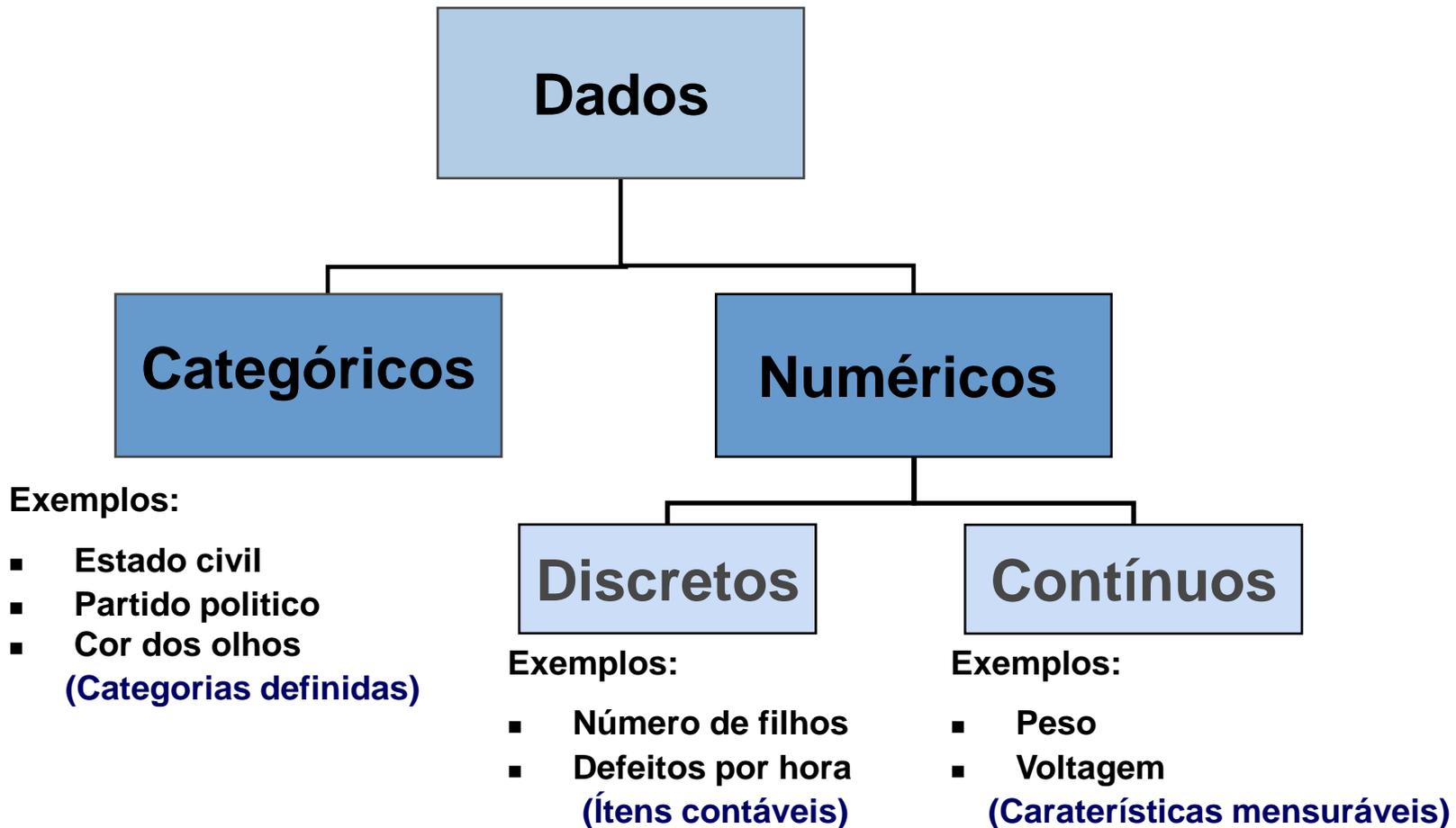
- **ser representativos**
 - Cuidado com amostras em que as pessoas escolhem se respondem ou não (particularmente importante em ciências sociais), podem não representar a população.
- conter informação suficiente
 - Dificilmente podemos extrair muitas informações de uma amostra pequena (ex: 2 pessoas).
- Método usado com frequência é a **seleção aleatória** de indivíduos a incluir na pesquisa (sorteio).



Tipos de Variáveis

- **Catagóricas** (qualitativas) variáveis cujos valores só podem ser colocado em categorias, tais como “sim” e “não”, partido político preferido, ...
- **Numéricas** (quantitativas) variáveis cujos valores representam quantidades, tais como peso, renda, número de filhos, número de máquinas, ...

Tipos de Variáveis



Dados e Variáveis

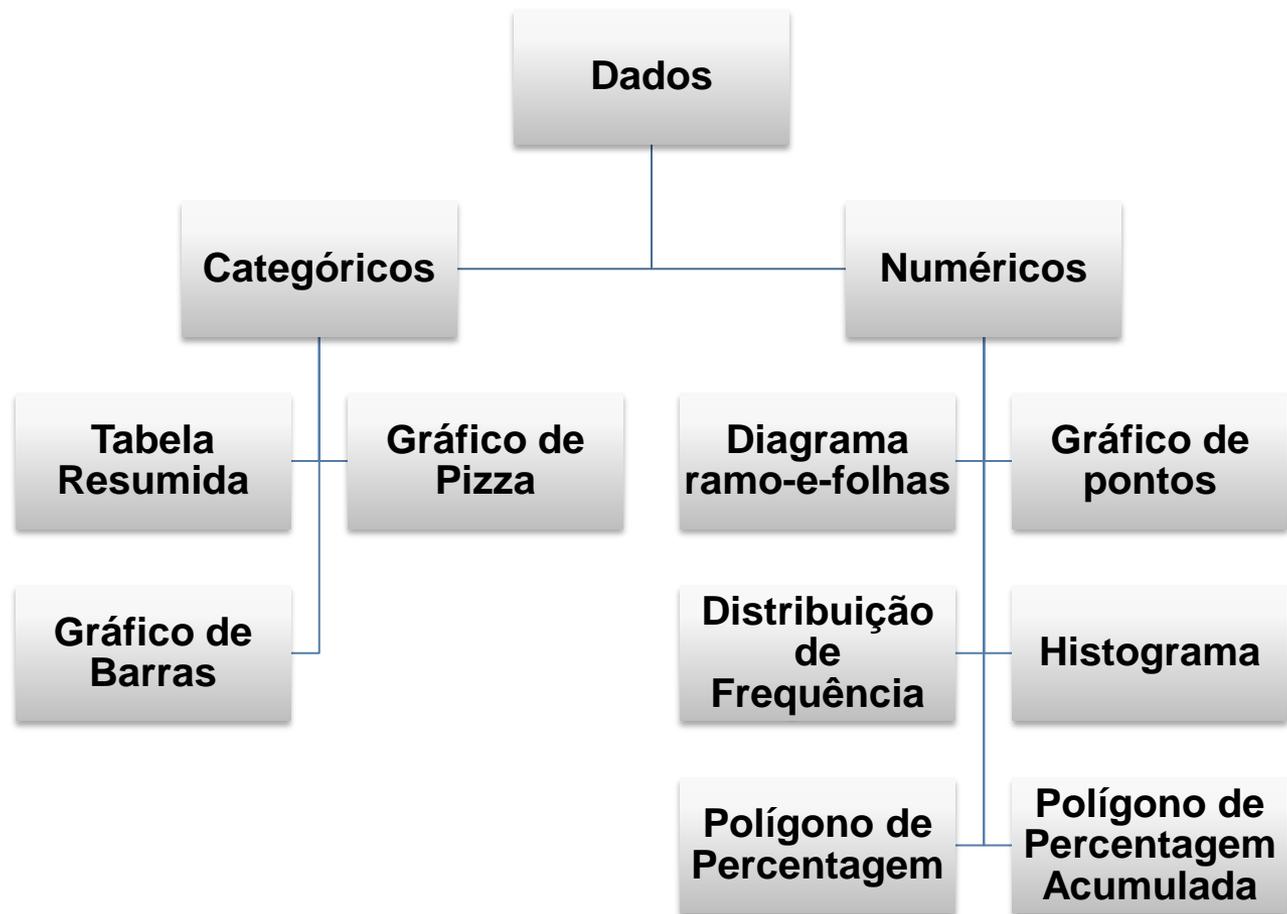


Exercício: O conjunto de dados abaixo descreve o consumo de combustível (em km/litro) para alguns modelos de sedãs médios.

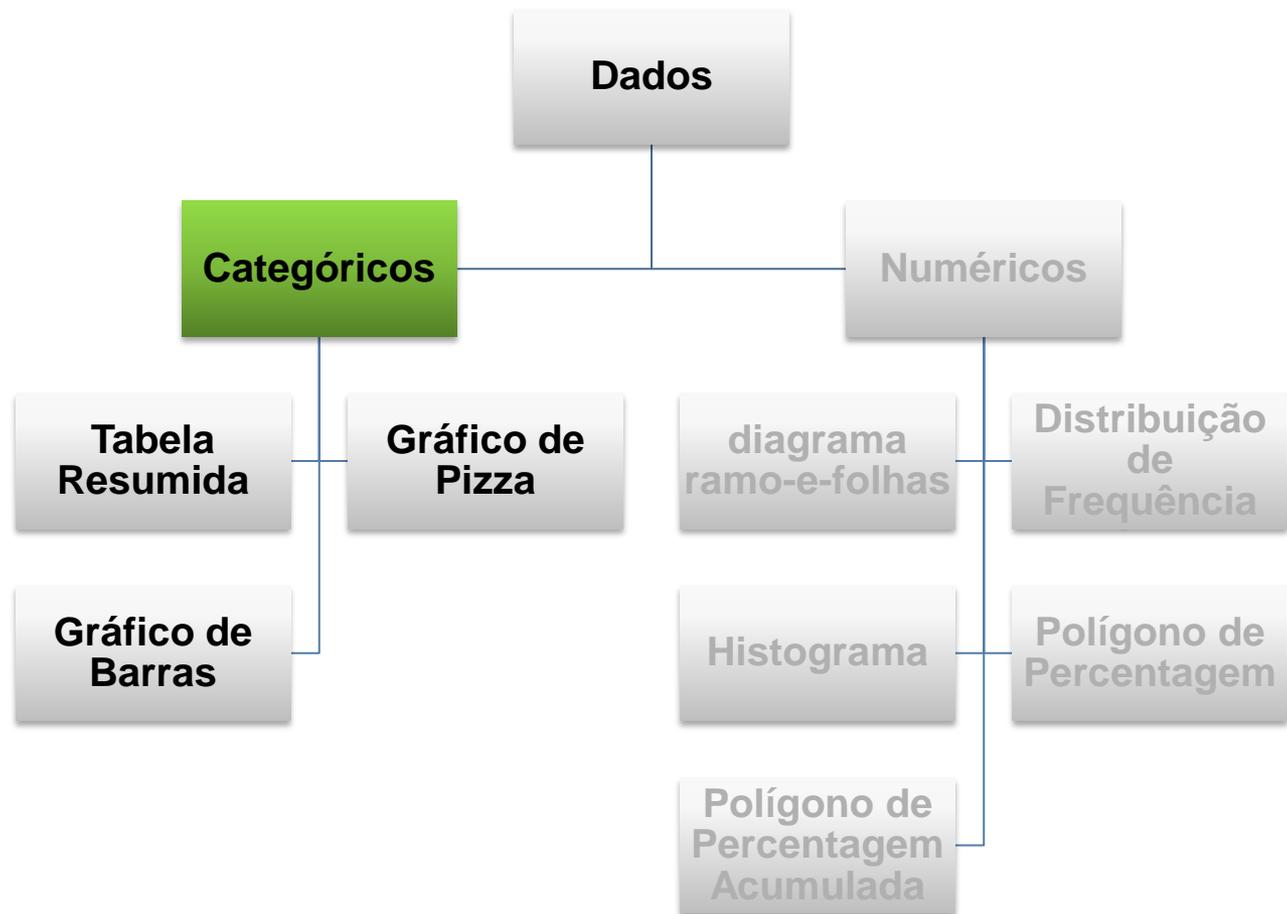
Modelo	Montadora	Direção	N° de Cilindros	Consumo Estrada (km/l)	Consumo Cidade (km/l)
...					
Marea	Fiat	Hidráulica	4	11	8.2
Focus	Ford	Hidráulica	4	12.8	9.7
Mégane	Renault	Elétrica	4	10.1	7.1
Corolla	Toyota	Hidráulica	4	13.8	10.4
...					

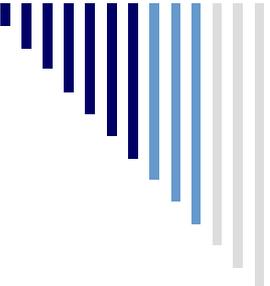
- Quais são os indivíduos nesse conjunto de dados?
- Quais variáveis são fornecidas? Quais são os tipos destas variáveis?
- Forneça uma amostra de tamanho 2 deste conjunto de dados.

Tipos de Gráficos e Tabelas



Tipos de Gráficos e Tabelas





Organizando Dados Categóricos: tabela resumida

- Uma **tabela resumida** ou distribuição de uma variável nos diz quais os valores assumidos por uma variável e qual a frequência (percentual ou absoluta) com que ela os assume.

Como você vai passar as festas de final de ano?	Frequência Absoluta
Em casa com a família	90
Viajando para visitar a família	76
Turismo	10
Colocando o trabalho em dia	10
Outros	14
Total	200

Como você vai passar as festas de final de ano?	Percentual
Em casa com a família	45%
Viajando para visitar a família	38%
Turismo	5%
Colocando o trabalho em dia	5%
Outros	7%
Total	100%

Organizando Dados Categóricos: tabela resumida



- Exercício: 20 alunos selecionados aleatoriamente de uma turma de Probabilidade e Estatística indicaram a ênfase que eles pretendem seguir. As respostas estão listadas abaixo. Construa uma tabela resumida para estes dados.

Comp	Mec	Comp	Amb	Amb
Comp	Mec	Mec	Comp	Telecom
Amb	Mec	Comp	Comp	Mec
Neuro	Mec	Comp	Amb	Mec

Organizando Dados Categóricos: tabela resumida

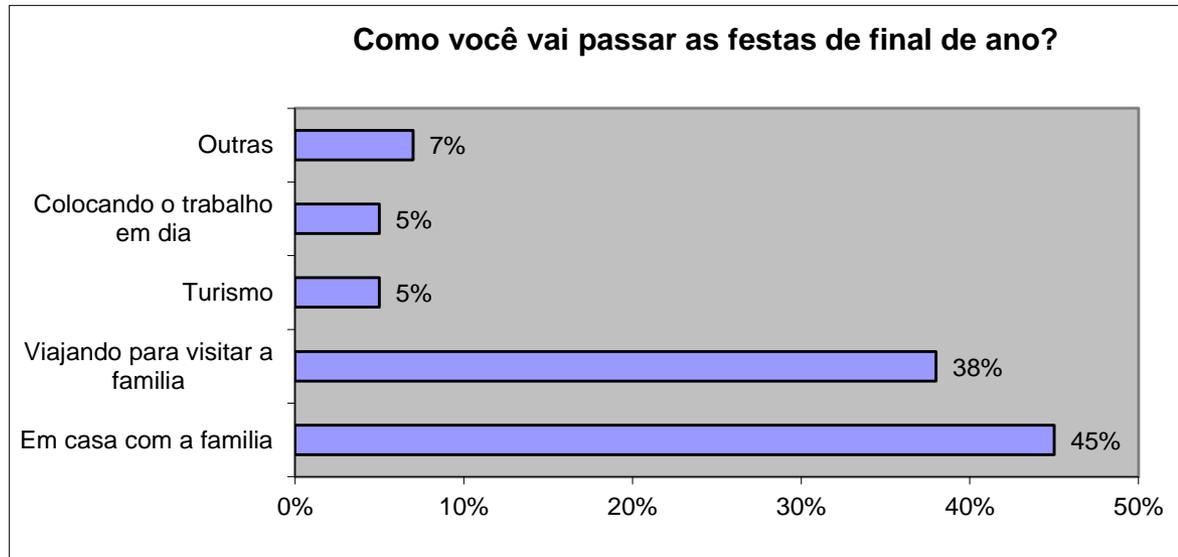


Solução:

Ênfase	Frequência absoluta	Frequência relativa
Ambiental	4	0.2
Computação	7	0.35
Mecânica	7	0.35
Neurociências	1	0.05
Telecomunicações	1	0.05

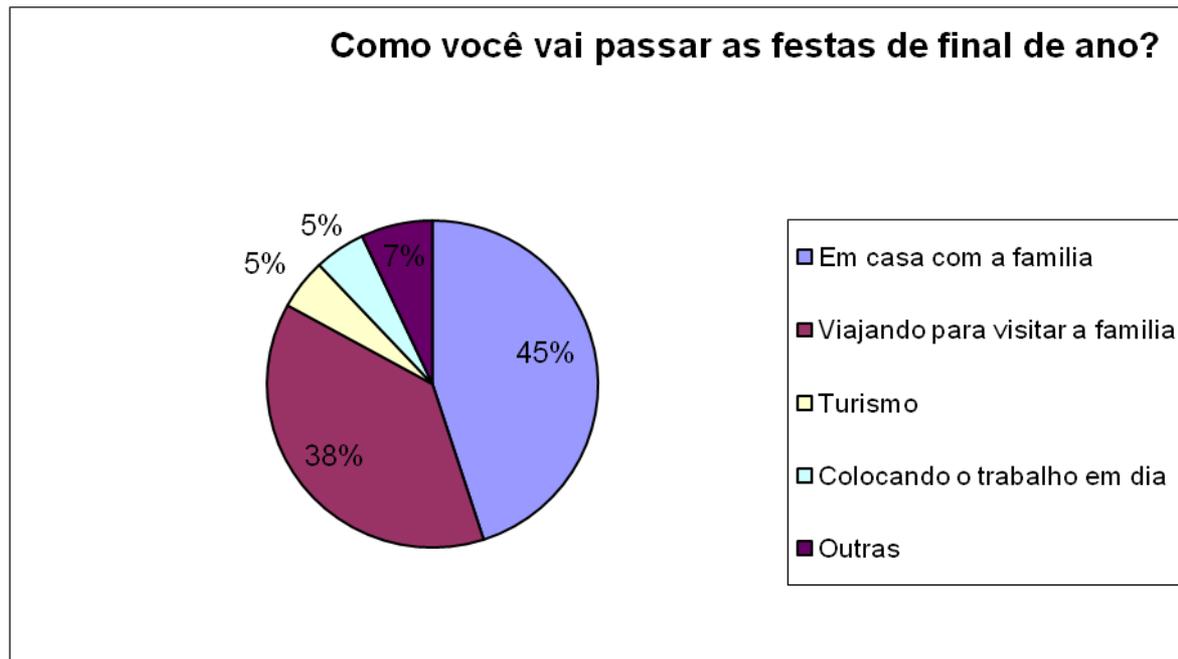
Organizando Dados Categóricos: gráfico de barras

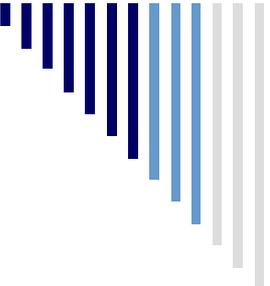
- Em um **gráfico de barras**, uma barra ilustra cada uma das categorias, cujo comprimento representa a *quantidade*, a *frequência* ou o *percentual* de valores que se posicionam em uma determinada categoria.



Organizando Dados Categóricos: o gráfico de pizza

- O **gráfico de pizza** é um círculo dividido em fatias que representam as categorias. O tamanho de cada fatia varia de acordo com a categoria.

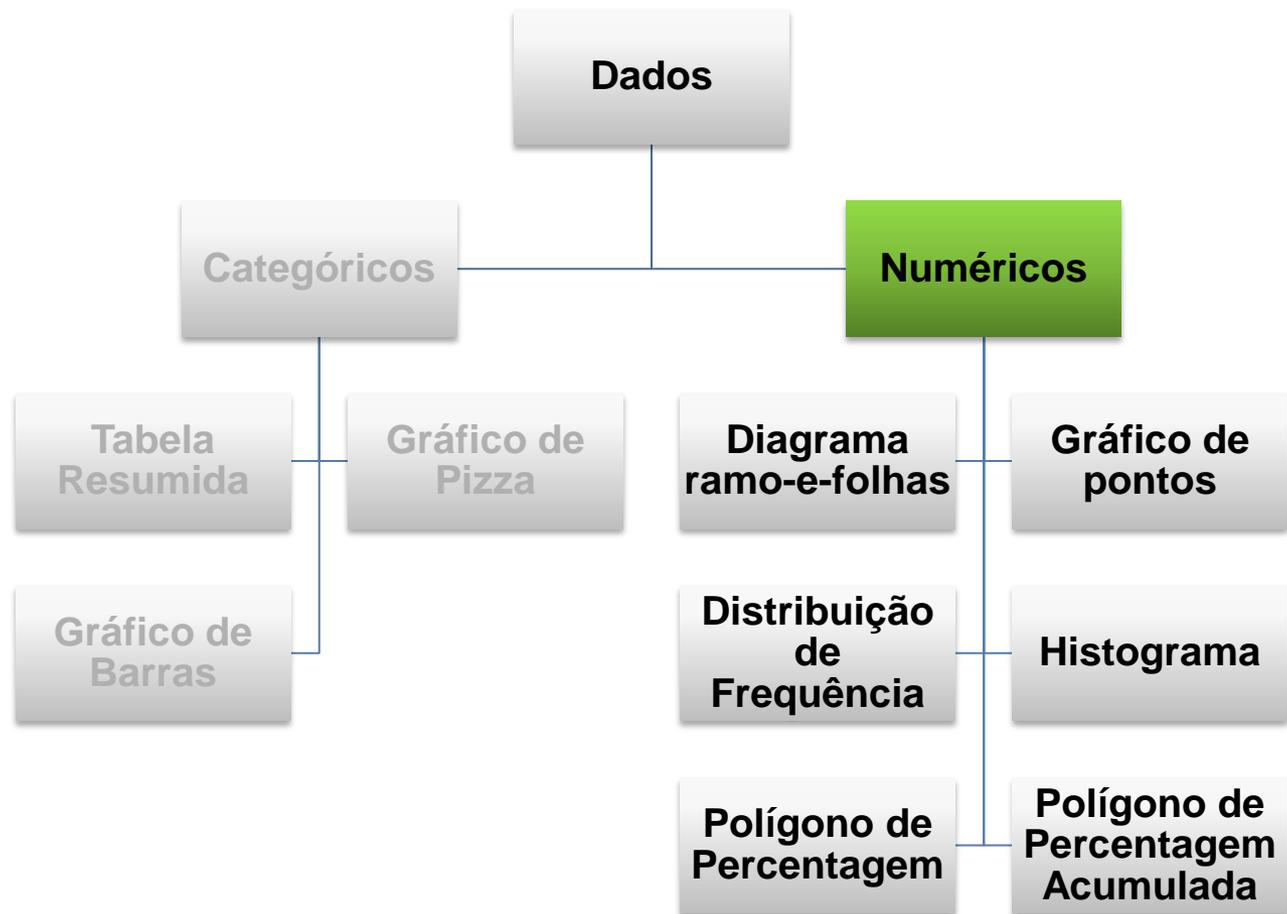


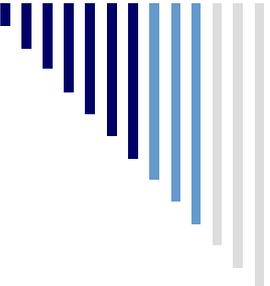


Dicas

- gráfico de pizza:
 - Enfatizar a relevância de uma categoria em relação ao total.
 - Se as categorias não somam 100%, criar categoria extra para o restante.
- Gráfico de barras:
 - Pode comparar qualquer conjunto de quantidades medidas na mesma unidade.
 - Mostrar categorias em ordem decrescente (ou crescente) ajuda a comparação visual entre categorias.

Tipos de Gráficos e Tabelas

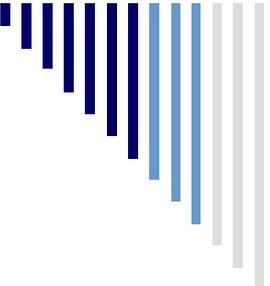




Organizando Dados Numéricos: Rol

- Um **rol** é uma sequência de dados ordenados de forma crescente ou decrescente.

Pesquisa com idade de estudantes universitários	Turma diurna					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Turma noturna					
	18	18	19	19	20	21
	23	28	32	33	39	52



Organizando Dados numéricos: diagrama ramo-e-folhas

- Uma **diagrama ramo-e-folhas** organiza dados em grupos (chamados de ramos ou caules) de tal modo que os valores em cada grupo (as folhas) se ramifiquem para a direita de cada linha.

Exemplo:

Variável: Idade dos alunos da turma diuna.

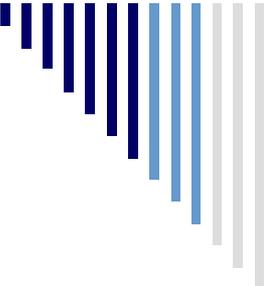
Dados em rol: 16, 17, 17, 18, 18, 18, 19, 19, 20, 20, 21, 22, 22, 25, 27, 32, 38, 42

Unidades de análise:

Ramo: dezenas

Folhas: unidades

Ramo	Folha
1	67788899
2	0012257
3	28
4	2



Organizando Dados numéricos: diagrama ramo-e-folhas

Para construir um diagrama ramo-e-folhas:

- 1) separe cada observação (número) em:
 - Um ramo: composto de todos os dígitos menos o final; e
 - Uma folha: composta pelo dígito final.
- 2) Escreva o ramo na coluna vertical (do menor para o maior)
sem pular nenhum número
- 3) Escreva cada folha à direita do ramo correspondente.



Organizando Dados numéricos: diagrama ramo-e-folhas

- Exemplo: Construa um diagrama ramo-e-folhas para a idade dos estudantes universitários da turma diurna:

16, 17, 17, 18, 18, 19, 19, 20, 20, 21, 22, 22, 25, 27, 32, 38, 42

Turma Diurna

Ramo	Folha
1	67788899
2	0012257
3	28
4	2

Unidades de análise:

Ramo: dezenas

Folhas: unidades

Organizando Dados numéricos: diagrama ramo-e-folhas



- Exercício: Construa um diagrama ramo-e-folhas para a idade dos estudantes universitários da turma noturna e compare com o da turma diurna.

18, 18, 19, 19, 20, 21, 23, 28, 32, 33, 39, 52

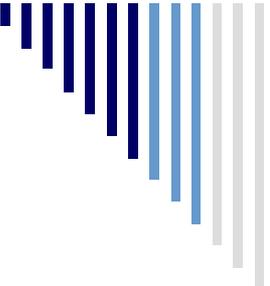
Turma Diurna

Ramo	Folhas
1	67788899
2	0012257
3	28
4	2

Turma Noturna

Ramo	Folhas
1	8899
2	0138
3	23
4	
5	2

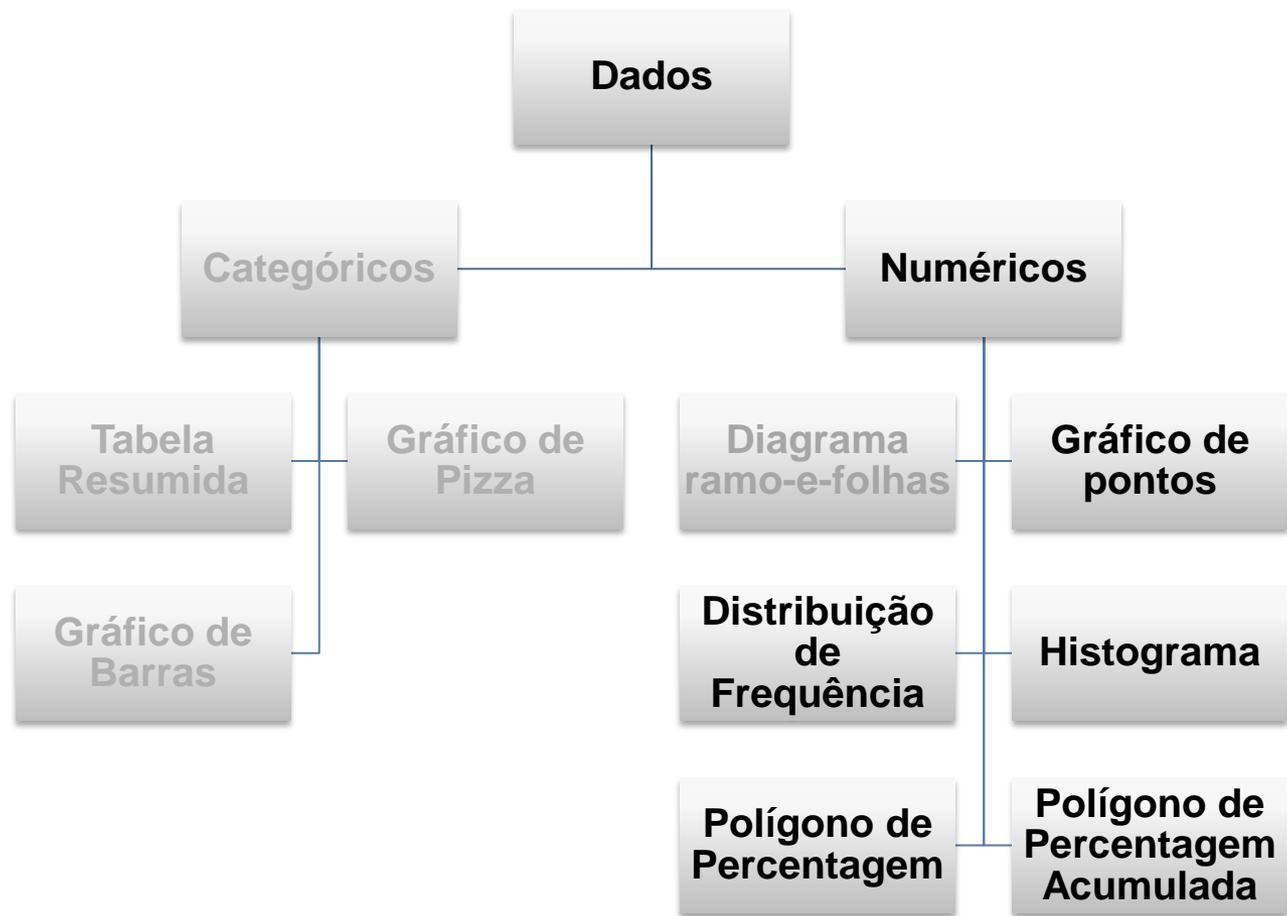
Unidades de análise:
Ramo: dezenas
Folhas: unidades

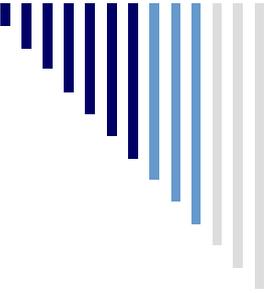


Analizando o Digrama ramo-e-folhas

- Um diagrama ramo-e-folhas transmite informações sobre:
 - Identificação de um **valor típico** ou representativo;
 - Extensão **dispersão** (afastamento) em relação ao valor típico;
 - Presença de **lacunas** nos dados;
 - Extensão da **simetria** da distribuição de valores;
 - Número e localização dos **picos** (máximos);
 - Presença de **valores fora da curva** (muito afastados dos demais valores).

Tipos de Gráficos e Tabelas





Organizando Dados numéricos: gráfico de pontos

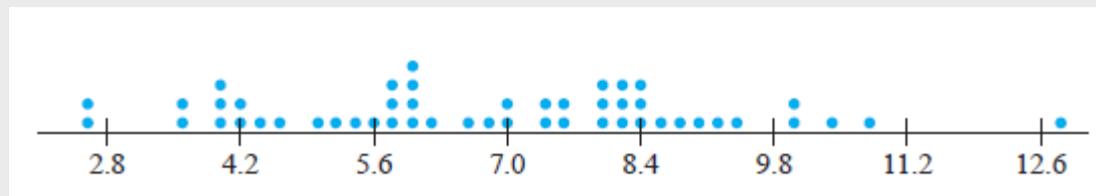
- Um gráfico bastante útil quando o conjunto de dados é pequeno é o gráfico de pontos.
- Nele, cada observação será representada por um ponto sobre o local correspondente em uma eixo.
- Quando um valor ocorre mais de uma vez, há um ponto para cada ocorrência e eles são empilhados verticalmente.

Organizando Dados numéricos: gráfico de pontos

- **Exemplo:** % das receitas dos estados aplicadas a universidade locais nos Estados Unidos (50 estados em ordem alfabética)

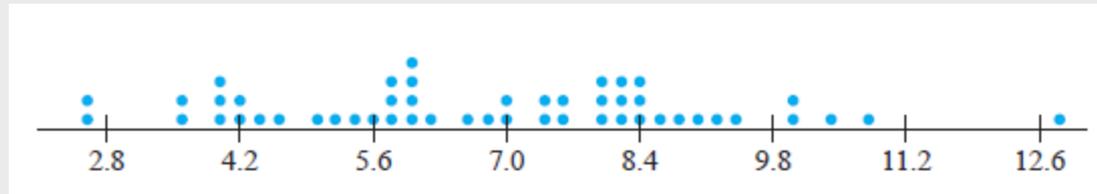
10.8	6.9	8.0	8.8	7.3	3.6	4.1	6.0	4.4	8.3
8.1	8.0	5.9	5.9	7.6	8.9	8.5	8.1	4.2	5.7
4.0	6.7	5.8	9.9	5.6	5.8	9.3	6.2	2.5	4.5
12.8	3.5	10.0	9.1	5.0	8.1	5.3	3.9	4.0	8.0
7.4	7.5	8.4	8.3	2.6	5.1	6.0	7.0	6.5	10.3

- Gráfico de pontos:



Organizando Dados numéricos: gráfico de pontos

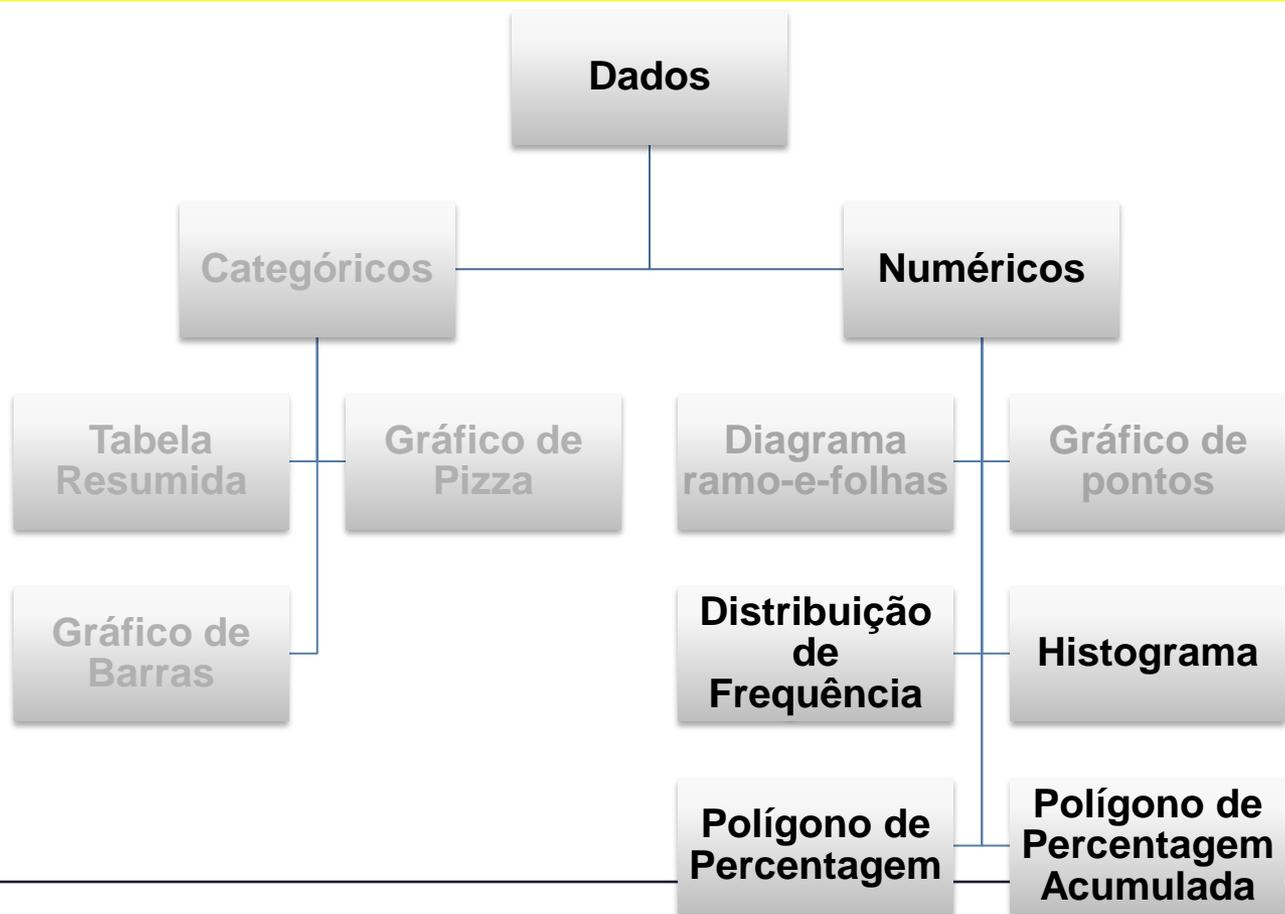
- **Exemplo:** % das receitas dos estados aplicadas a universidade locais nos Estados Unidos (50 estados em ordem alfabética)
- Gráfico de pontos:



- Podemos observar os mesmo aspectos que para o diagrama ramo-e-folhas:
 - Valor típico, dispersão, extremos, simetria e lacunas.
 - **O que vocês observam nesse gráfico?**

Tipos de Gráficos e Tabelas

E se o número de dados na base for muito grande?? Ramo-e-folhas não é conveniente...





Organizando Dados Numéricos: distribuição de frequência

- A **distribuição de frequência** é uma *tabela resumida* na qual os dados são dispostos em grupos de classe ordenados numericamente.
- Atenção para selecionar o número apropriado de **grupos de classe** para a tabela: escolher a *amplitude* adequada para cada grupo de classe e os *limites* de cada grupo de classe, evitando sobreposições.
- Para determinar a **amplitude de um intervalo de classe**:
 - Calcule: $\frac{\text{maior valor} - \text{menor valor}}{n^\circ \text{ de classes}}$
 - Escolha *um inteiro próximo* ao valor calculado



Organizando Dados Numéricos: distribuição de frequência

Exemplo: Um fabricante de isolante térmico selecionou aleatoriamente 20 dias no inverno e registrou a temperatura máxima do dia (em °F)

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Em rol:

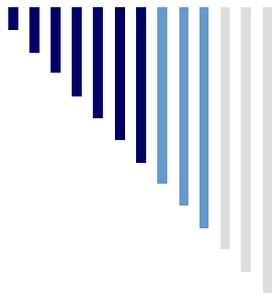
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Para ter uma ideia melhor de como variam as temperaturas máximas, vamos construir uma **tabela com distribuição de frequência** dos dados! Que **classes** vamos usar?



Organizando Dados Numéricos: distribuição de frequência

- Reorganizar os dados em ordem crescente:
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Encontrar a amplitude: **$58 - 12 = 46$**
- Escolher o número de grupos: **5 (geralmente entre 5 e 15)**
- Calcular o intervalo de classe (amplitude): **10 (aprox. $46/5$)**
- Determinar os limites das classes: **10, 20, 30, 40, 50, 60**
- Contar observações e designar a classes



Organizando Dados Numéricos: distribuição de frequência

Classe	Frequência	Frequência Relativa	Porcentagem (%)
[10,20)	3	0.15	15
[20,30)	6	0.30	30
[30,40)	5	0.25	25
[40,50)	4	0.20	20
[50,60)	2	0.10	10
Total	20	1.00	100

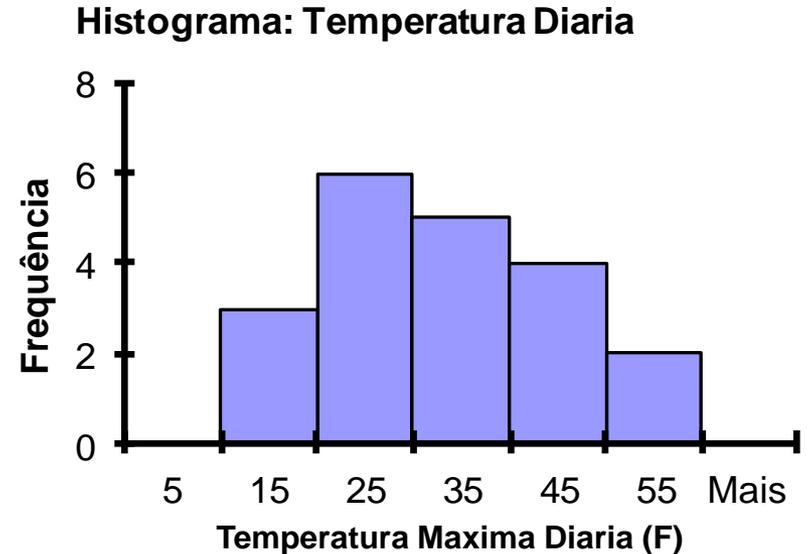
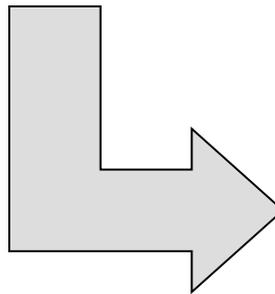


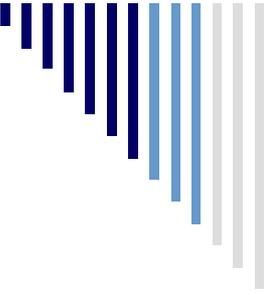
Organizando Dados Numéricos: Histograma

- Um gráfico com a distribuição de frequência de dados é chamado de **histograma**.
- Os **limites de classe** (ou **pontos médios** de classes) são mostrados no eixo horizontal.
- O eixo vertical representa a **frequência**, **frequência relativa** ou **percentagem**.
- Barras de tamanho apropriado são usadas para representar o número de observações em cada classe.

Organizando Dados Numéricos: Histograma

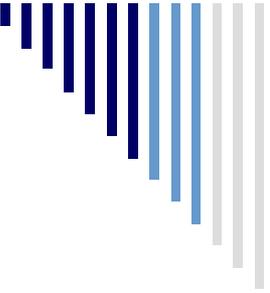
Classe	Frequência	Frequência Relativa	Porcentagem
[10, 20)	3	.15	15
[20, 30)	6	.30	30
[30, 40)	5	.25	25
[40, 50)	4	.20	20
[50, 60]	2	.10	10
Total	20	1.00	100





Organizando Dados Numéricos: Histograma

- Quando for analisar um histograma observe:
 - Um padrão geral: forma, centro e dispersão dos dados
 - Valores atípicos: valor individual que está fora do padrão geral
- Quantas classes usar?
 - A definição de classes pode alterar bastante o histograma.
 - Os programas computacionais escolhem as classes automaticamente usando a fórmula de sturges. Mas, se você quiser, sempre pode escolher as classes.



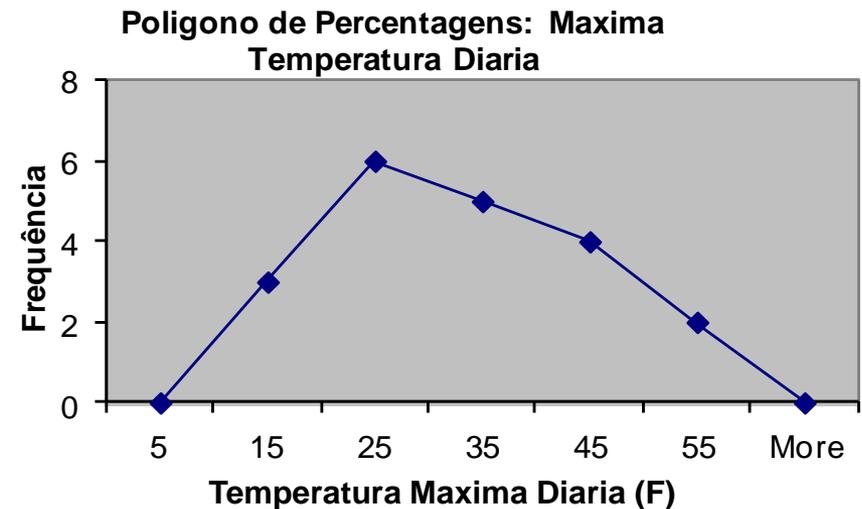
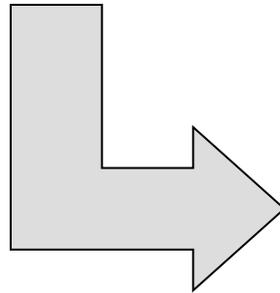
Organizando Dados Numéricos: polígono de percentagem

- Um **polígono de percentagem** é formado fazendo-se com que o ponto médio de cada classe represente os dados naquela classe e, depois, interligando-se a sequência de pontos médios em suas respectivas percentagens de classe.
- Um **polígono de percentagens acumuladas (Ogivas)** exhibe a variável de interesse ao longo do eixo X e a *percentagem acumulada* ao longo do eixo Y .

Organizando Dados Numéricos: polígono de percentagem

Classe	Frequência	Frequência Relativa	Percentagem
[10, 20)	3	.15	15
[20, 30)	6	.30	30
[30, 40)	5	.25	25
[40, 50)	4	.20	20
[50, 60]	2	.10	10
Total	20	1.00	100

(Em um polígono de percentagens o eixo vertical mostra a percentagem de todas as observações por classe)

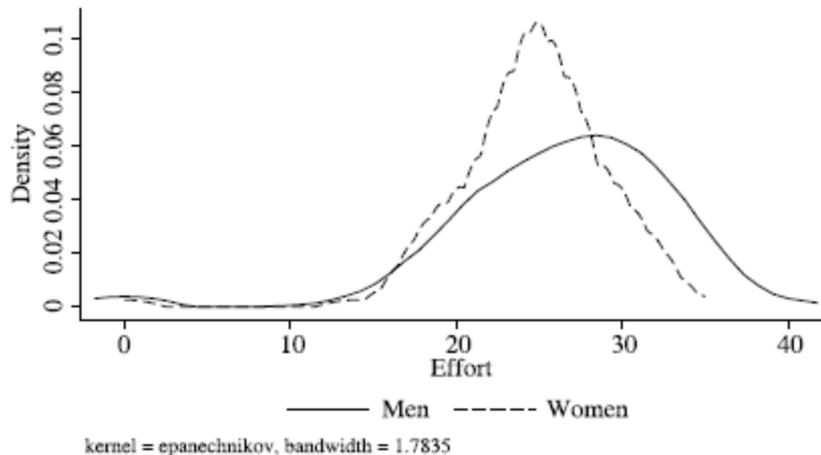


Organizando Dados numéricos: polígono de percentagem

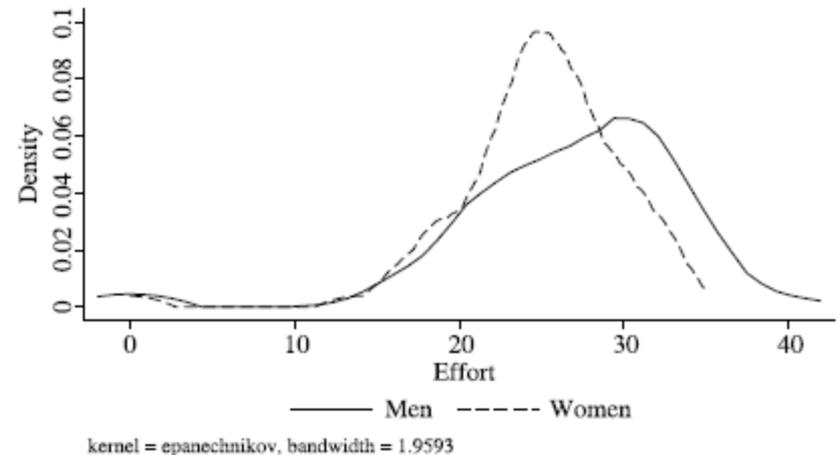
Polígonos de percentagem, em sua versão suavizada, são comumente usados em artigos científicos para fazer comparações entre grupos distintos ou ao longo do tempo.

358 Gill and Prowse

Quantitative Economics 5 (2014)



(a) Distributions of efforts for rounds 3–10

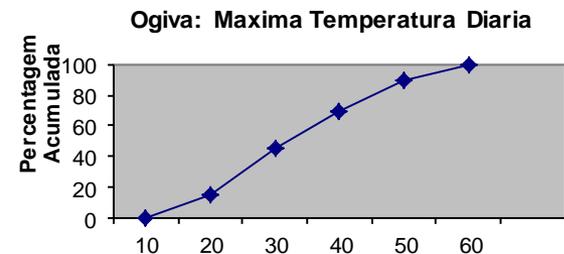
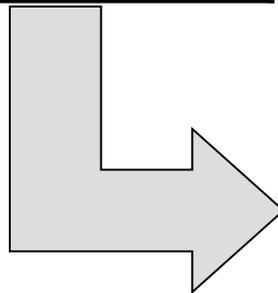


(b) Distributions of efforts for rounds 6–10

FIGURE 2. Distributions of effort choices.

Organizando Dados Numéricos: polígono de percentagem acumulada

Classe	Limite inferior	% Menor do que o Limite inferior
[10,20)	10	0
[20,30)	20	15
[30,40)	30	45
[40,50)	40	70
[50,60)	50	90
	60	100



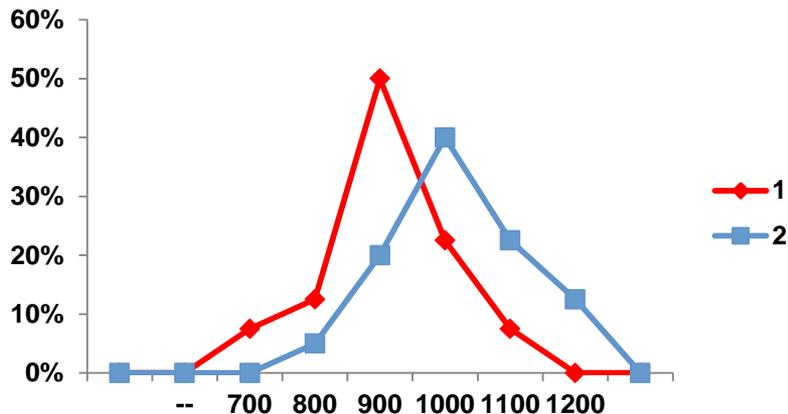
Temperatura Maxima Diaria (F)

Organizando Dados Numéricos: polígono de percentagem acumulada

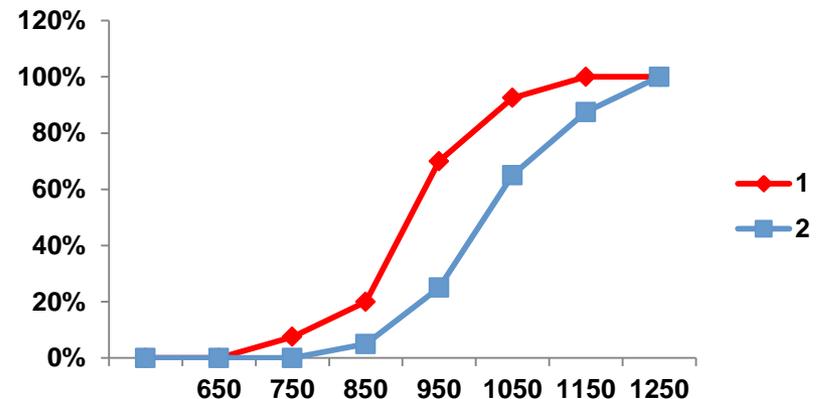
Polígonos de percentagem acumulada também são usados para fazer comparações entre grupos ou entre métodos.

Exemplo: comparando vida útil de 2 fabricantes de lâmpadas.

Polígono de percentagem

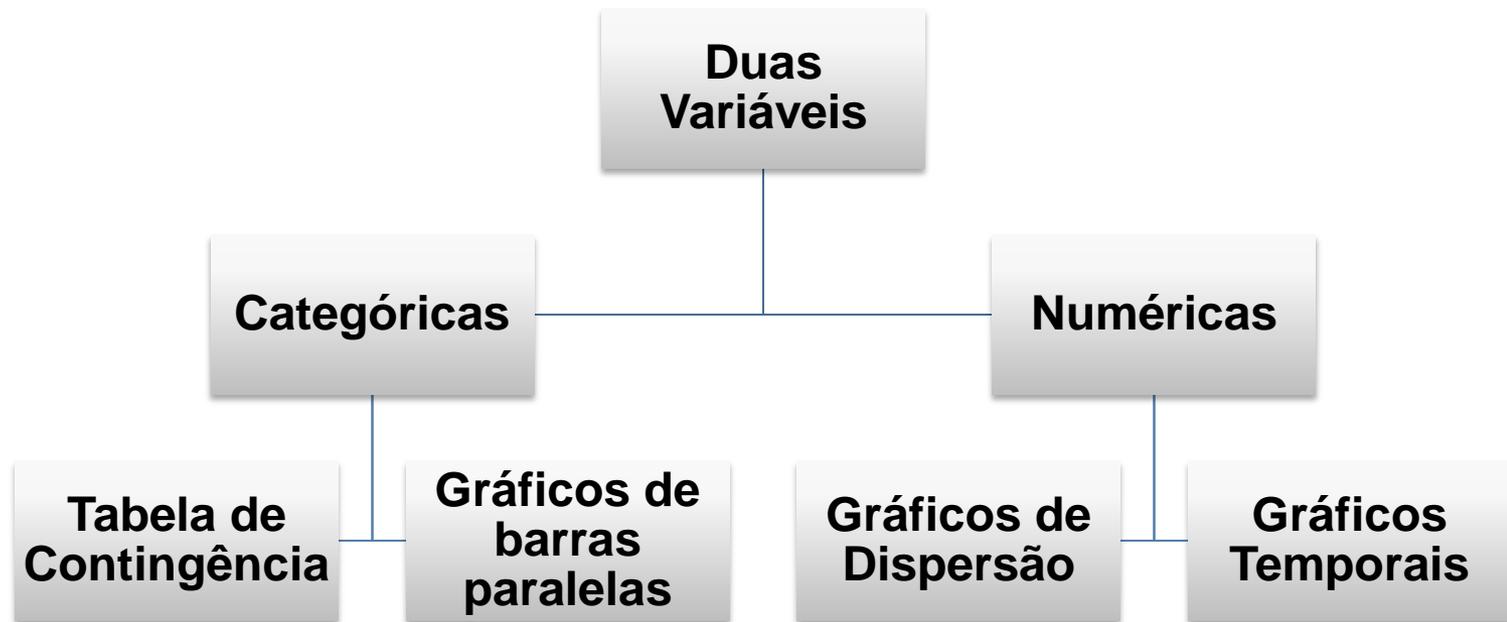


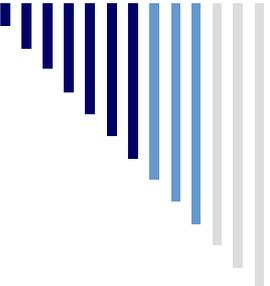
Políg. perc. Acumulada



Gráficos e Tabelas para Duas Variáveis

Até agora analisamos apenas uma variável separadamente. E se quisermos analisar a relação entre duas variáveis?





Tabulações Cruzadas: tabela de contingência

- Uma **tabela de contingência** apresenta os resultados de duas variáveis categóricas. As respostas combinadas são classificadas de tal modo que as categorias de uma variável fiquem localizadas nas linhas enquanto as categorias da outra variável ficam localizadas nas colunas.
- Os valores localizados nas intersecções de linhas e colunas são chamados de **células**. Células podem conter:
 - Frequências absolutas
 - Percentagens
 - Percentagens por linha ou por coluna.



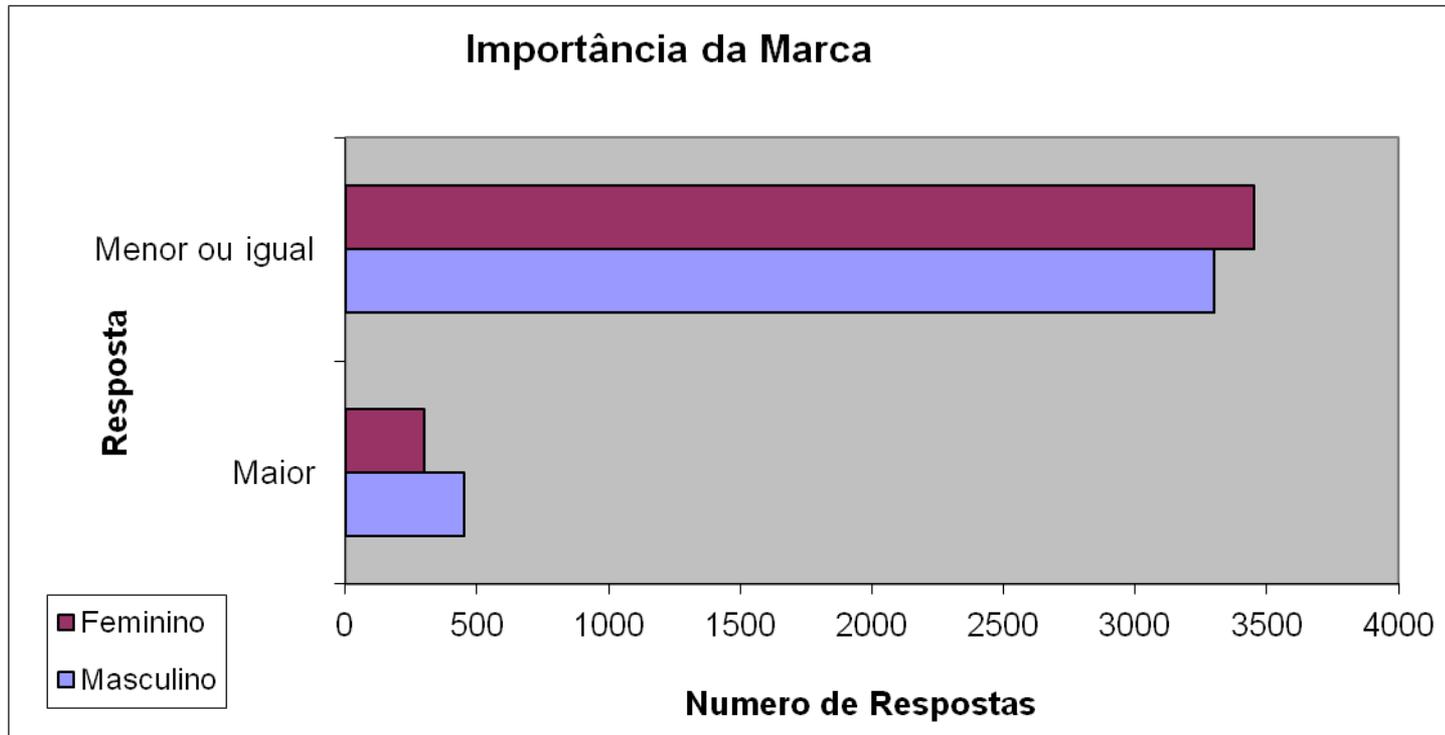
Tabulações Cruzadas: tabela de contingência

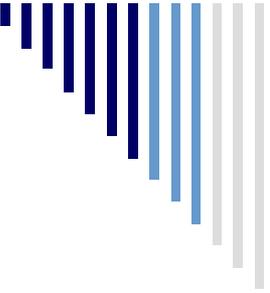
Exemplo: Uma pesquisa foi realizada para determinar a **importância de uma marca para consumidores em comparação a anos anteriores**. Os resultados por gênero são apresentados abaixo:

	Sexo		
Importância da marca	Masculino	Feminino	Total
Maior	450	300	750
Menor ou igual	3300	3450	6750
Total	3750	3750	7500

Quais são as variáveis? Quais são as categorias de cada variável?

Tabulações Cruzadas: gráfico de barras paralelas





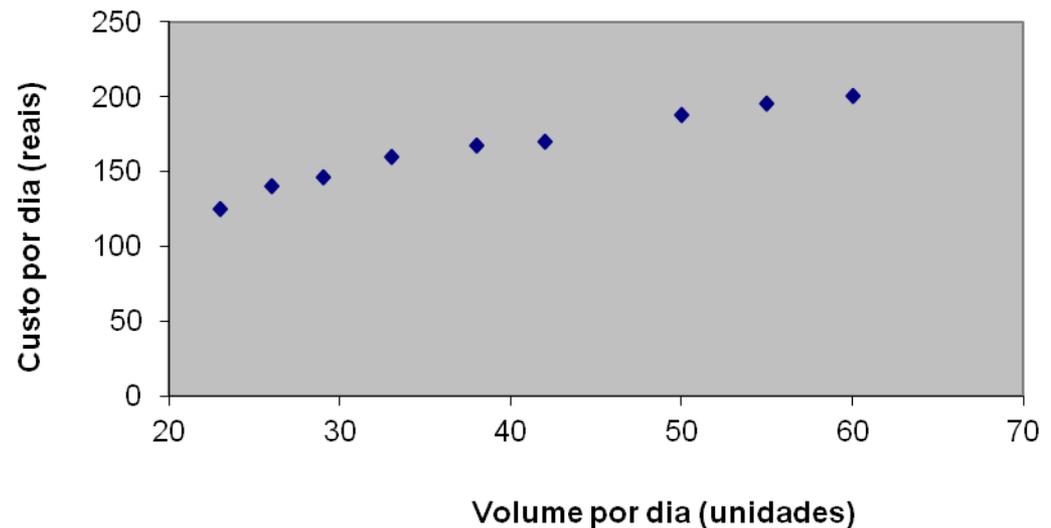
Gráficos de Dispersão

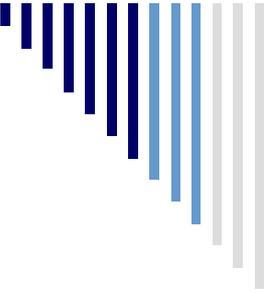
- **Gráficos de Dispersão** são usados examinar possíveis relações entre duas variáveis numéricas.
- Uma variável é medida no eixo vertical e a outra no eixo horizontal.

Gráfico de Dispersão

Volume por dia	Custo por dia
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

Custo por dia vs. volume por dia





Séries Temporais

- Um **gráfico de séries temporais** é usado para estudar padrões nos valores de uma variável numérica ao longo do tempo. Cada valor é inserido no gráfico como um ponto de duas dimensões com o período de tempo no eixo X e a variável de interesse no eixo Y.



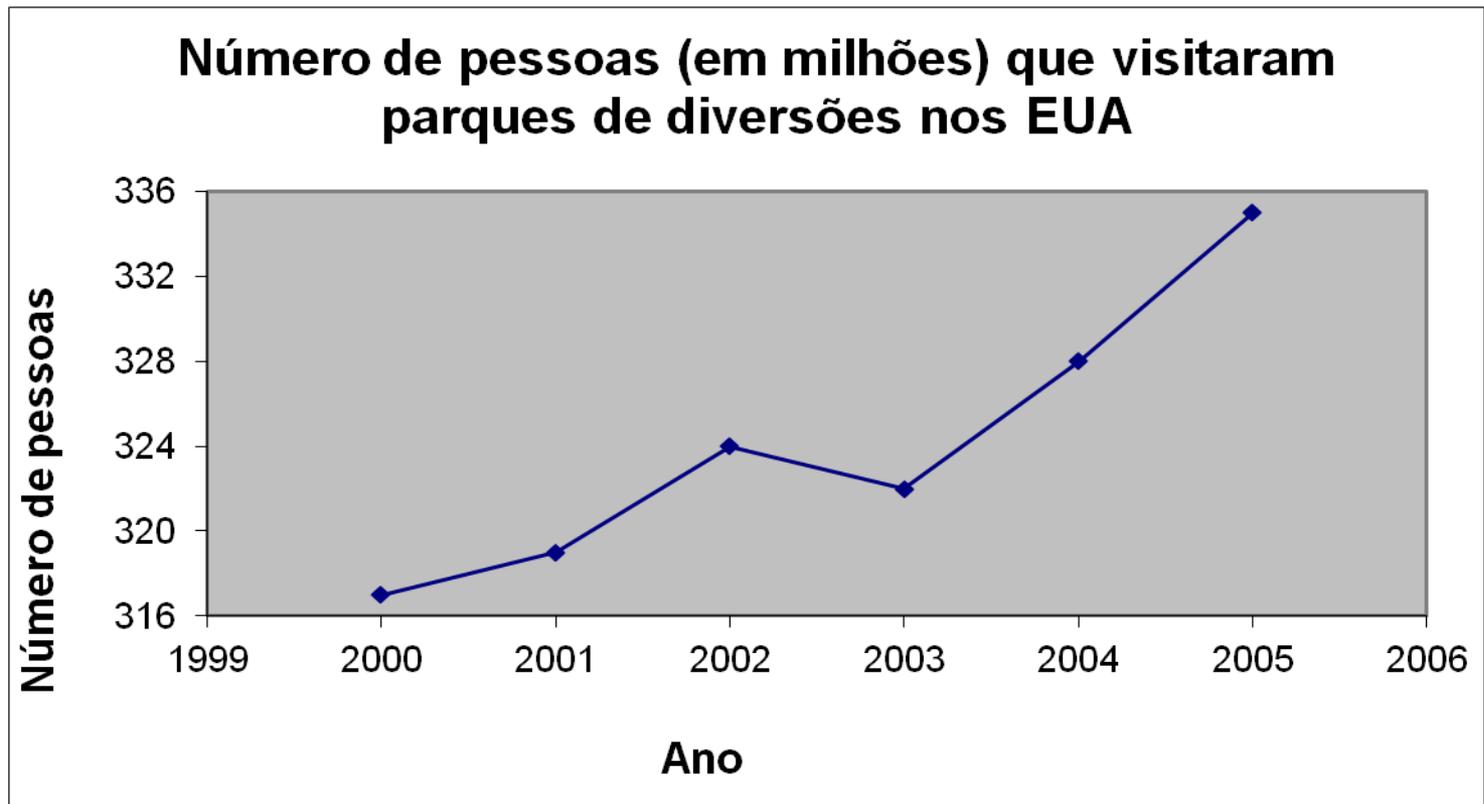
Exemplo de Séries Temporais

Número de pessoas (em milhões) que visitaram um parque de diversões nos EUA de 2000 a 2005.

Ano	Número de pessoas
2000	317
2001	319
2002	324
2003	322
2004	328
2005	335

Represente graficamente a evolução do número de visitantes ao parque de diversões nos anos indicados.

Exemplo de Séries Temporais





Princípios de Gráficos

- O gráfico não deve distorcer os dados.
- O gráfico não deve conter adornos desnecessários (também conhecidos como sucata de gráficos).
- A escala do eixo vertical deve sempre começar no zero.
- Todos os eixos devem ser nomeados.
- Um gráfico deve conter um título.
- O gráfico mais simples possível deve ser utilizado para um dado conjunto de dados.



Resumo

Nesta parte, nos vimos:

- Como organizar **dados categóricos** usando tabela resumida, gráfico de barras, gráfico de pizza.
- Como organizar **dados numéricos** usando o diagrama ramo-e-folhas, o gráfico de pontos, a distribuição de frequências, o histograma e os polígonos de percentagem e de percentagem acumulada.
- Como examinar **tabulações cruzadas** usando tabela de contingência e gráficos de barras paralelas.



Resumo

Nesta parte, nos vimos:

- Como utilizar gráfico de dispersão e gráfico de séries temporais.
- O que fazer e o que não fazer para ter gráficos com boa apresentação.

- (Devore, ex. 11) Os dados a seguir apresentam notas de alunos em uma determinada prova. Para facilitar, as notas já estão em ordem crescente.

60	63	64	66	66	67	68	69	69	70
70	71	72	72	72	74	74	80	80	81
81	81	81	81	82	82	83	84	84	85
85	85	87	88	89	89	90	93	95	98

- Construa um diagrama ramo-e-folhas para estes dados.
- Que características dos dados são observadas através do diagrama ramo-e-folhas?

- (Devore, ex. 11) Os dados a seguir apresentam notas de alunos em uma determinada prova. Para facilitar, as notas já estão em ordem crescente.

60	63	64	66	66	67	68	69	69	70
70	71	72	72	72	74	74	80	80	81
81	81	81	81	82	82	83	84	84	85
85	85	87	88	89	89	90	93	95	98

- O diagrama ramo-e-folhas anterior com apenas 4 ramos (6, 7, 8, e 9) não fornece uma descrição detalhada das notas. Nesse tipo de situação, podemos usar ramos repetidos. Por exemplo, podemos repetir o ramo 6 duas vezes, usando "6-" para as notas mais baixas entre 60 e 70 (folhas 0, 1, 2, 3 e 4) e "6+" para as notas mais altas entre 60 e 70 (folhas 5, 6, 7, 8 e 9). De forma semelhante, outros ramos podem ser repetidos para construir um diagrama com 8 linhas.
 - Construa este diagrama. Que características dos dados são observadas através do diagrama ramo-e-folhas?